



中国计算机学会学术著作丛书

知识发现 (第二版)

Knowledge Discovery
(Second Edition)

史忠植 著

清华大学出版社



中国计算机学会学术著作丛书

知识发现 (第二版)

Knowledge Discovery
(Second Edition)

史忠植 著

清华大学出版社
北京

内 容 简 介

知识发现是从数据集中识别出有效的、新颖的、潜在有用的,以及最终可理解的模式的非平凡过程。知识发现将信息变为知识,从数据资源中发现知识宝藏,将为知识创新和知识经济的发展作出贡献。

本书全面而又系统地介绍了知识发现的方法和技术,反映了当前知识发现研究的最新成果和进展。全书共分 15 章。第 1 章是绪论,概述知识发现的重要概念和发展过程。下面三章重点讨论分类问题,包括决策树、支持向量机和迁移学习。第 5 章阐述聚类分析。第 6 章是关联规则。第 7 章讨论粗糙集和粒度计算。第 8 章介绍神经网络,书中着重介绍几种实用的算法。第 9 章探讨贝叶斯网络。第 10 章讨论隐马尔可夫模型。第 11 章探讨图挖掘。第 12 章讨论进化计算和遗传算法。第 13 章探讨分布式知识发现,它使海量数据挖掘成为可能。最后两章以 Web 知识发现、认知神经科学为例,介绍知识发现的应用。

本书内容新颖,认真总结了作者的科研成果,取材国内外最新资料,反映了当前该领域的研究水平。论述力求概念清晰,表达准确,算法丰富,突出理论联系实际,富有启发性。

本书可以用作高等院校有关专业的研究生和高年级本科生的知识发现、数据挖掘、机器学习等课程教材,也可供从事知识发现、数据挖掘、机器学习、智能信息处理、模式识别、智能控制研究和知识管理的科技人员阅读参考。

版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

知识发现/史忠植著.--2 版.--北京:清华大学出版社,2011.1

(中国计算机学会学术著作丛书)

ISBN 978-7-302-23957-4

I. ①知… II. ①史… III. ①知识工程 ②人工智能 IV. ①TP18

中国版本图书馆 CIP 数据核字(2010)第 199169 号

责任编辑:薛 慧

责任校对:刘玉霞

责任印制:王秀菊

出版发行:清华大学出版社

地 址:北京清华大学学研大厦 A 座

<http://www.tup.com.cn>

邮 编:100084

社 总 机:010-62770175

邮 购:010-62786544

投稿与读者服务:010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈:010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者:清华大学印刷厂

经 销:全国新华书店

开 本:175×245 印 张:32.25 字 数:625 千字

版 次:2011 年 1 月第 2 版 印 次:2011 年 1 月第 1 次印刷

印 数:1~3000

定 价:59.00 元

产品编号:035243-01

序

Preface

第

一台电子计算机诞生于 20 世纪 40 年代。到目前为止,计算机的发展已远远超出了其创始者的想象。计算机的处理能力越来越强,应用面越来越广,应用领域也从单纯的科学计算渗透到社会生活的方方面面:从工业、国防、医疗、教育、娱乐直至人们的日常生活,计算机的影响可谓无处不在。

计算机之所以能取得上述地位并成为全球最具活力的产业,原因在于其高速的计算能力、庞大的存储能力以及友好、灵活的用户界面。而这些新技术及其应用有赖于研究人员多年不懈的努力。学术研究是应用研究的基础,也是技术发展的动力。

自 1992 年起,清华大学出版社与广西科学技术出版社为促进我国计算机科学技术与产业的发展,推动计算机科技著作的出版,设立了“计算机学术著作出版基金”,并将资助出版的著作列为中国计算机学会的学术著作丛书。时至今日,本套丛书已出版学术专著近 50 种,产生了很好的社会影响,有的专著具有很高的学术水平,有的则奠定了一类学术研究的基础。中国计算机学会一直将学术著作的出版作为学会的一项主要工作。本届理事会将秉承这一传统,继续大力支持本套丛书的

出版,鼓励科技工作者写出更多的优秀学术著作,多出好书,多出精品,为提高我国的知识创新和技术创新能力,促进计算机科学技术的发展和进步作出更大的贡献。

中国计算机学会

2002年6月14日

前言

Foreword

知

识发现将从数据集中识别出有效的、新颖的、潜在有用的,以及最终可理解的模式的非平凡过程。它集数据收集、数据清洗、降维、规则归纳、模式识别、结果分析与评估、可视化输出等多种功能于一身,是统计学、机器学习、神经网络、计算机科学、模式识别、人工智能及其他学科相结合的产物。

从 20 世纪 80 年代中期的小范围研究到如今的蓬勃发展,知识发现已经在企业界与科学界占据了一席之地。在信息管理、查询响应、决策支持、过程控制、搜索引擎、基因分析等许多方面得到了广泛应用。帮助企业进行客户关系管理,优化生产过程,减少不必要的投资,提高资金周转和回报。帮助人们迅速获取所需的知识和信息,提高工作效率,改进服务质量。不同领域的人们都运用知识发现,将信息变为知识,从数据资源中发现知识宝藏。

本书第一版自 2002 年出版以来,国内外知识发现的研究者们取得了许多重要进展,涌现出了许多新概念、新方法、新算法,应用领域也日新月异。本书第二版对原书作了重大修改,增加迁移学习、聚类分析、图挖掘、分布式知识发现、认知神经科学知识发现等内容,尽可能反映这近十年间知识发现、数据挖掘,及其应用所取得的新成果,指

出当今活跃的研究方向和课题,给读者有所启发。

本书全面而又系统地介绍知识发现的方法和技术。全书共分 15 章。第 1 章是绪论,概述知识发现的重要概念和发展过程。第 2 章讨论决策树,它是归纳学习方法中最实用的一种技术,新版中增加决策树并行挖掘算法。广泛应用的支持向量机在第 3 章讨论,增加了改进算法。第 4 章探讨迁移学习。前面三章重点讨论分类问题。第 5 章阐述聚类分析。第 6 章是关联规则,它是近几年应用最为广泛的挖掘方法之一。第 7 章讨论粗糙集,增加了粒度计算的内容。第 8 章介绍神经网络,书中着重介绍几种实用的算法。第 9 章探讨贝叶斯网络,贝叶斯网络可以处理不完整和带有噪声的数据集,它用概率测度的权重来描述数据间的相关性。它在近几年知识发现研究中是极其活跃的研究课题。第 10 章讨论隐马尔可夫模型。第 11 章探讨图挖掘,这是近期极其活跃的研究领域,在互联网分析、社会计算、生物信息学等方面得到广泛应用。第 12 章讨论进化计算和遗传算法。第 13 章探讨分布式知识发现,它使海量数据挖掘成为可行,重点阐述多主体(multi-agent)、网格计算、云计算环境下的海量数据挖掘。最后以 Web 知识发现、认知神经科学为例,介绍知识发现的应用。第 14 章关于 Web 知识发现。第 15 章介绍认知神经科学中如何运用知识发现技术分析认知神经科学的实验结果,探索智能的本质和机理。

本书总结了作者和中国科学院计算技术研究所智能科学实验室多年的科研成果,也吸取了国内外同行的研究成果和有关文献的精华,在此谨向这些成果和文献的作者表示感谢,他们的丰硕成果和贡献是本书学术思想的重要源泉。本书的顺利撰写离不开中国科学院计算技术研究所智能科学实验室同事们的贡献和支持,特别是何清、叶世伟、李晓黎、叶施仁、宫秀军、刘少辉、郑毅、郑金华、胡宏、张建、王军、张颖、潘谦红、吴斌、贾自艳、秦亮曦、游湘涛、任力安、李宝东、梁吉业、丁世飞、蒙祖强、施智平、李清勇、郑征、罗平、罗杰文、石志伟、石川、李志清、李志欣、马慧芳、庄福振、刘曦、崔志华、周菁等给予了帮助和支持,在此一并表示感谢。

本书研究工作得到自然科学基金重点项目“基于云计算的海量数据挖掘”(批准号:61035003)、“基于感知学习和语言认知的智能计算模型研究”(批准号:60435010)、“Web 搜索与挖掘的新理论与方法”(批准号:60933004)、自然科学基金项目“语义 Web 服务的逻辑基础”(批准号:60775035)等的支持。感谢国家重点基础研究发展计划“基于视觉认知的非结构化信息处理理论与关键技术”(项目编号:2007CB311000)、国家 863 高技术探索项目“软件自治愈与自恢复技术”(项目编号:2007AA01Z132)等项目的支持。

本书内容新颖,反映了当前该领域国内外的研究水平。论述力求概念清晰,表达准确,算法丰富,突出理论联系实际,通过实例说明原理,富有启发性。

本书可以用作高等院校有关专业的研究生和高年级本科生的知识发现、数据挖掘、机器学习等课程教材,也可供从事知识发现、数据挖掘、机器学习、智能信息处理、模式识别、智能控制研究和知识管理的科技人员阅读参考。

由于作者水平有限,加上知识发现发展很快,研究领域广泛,书中不妥和错误之处在所难免,恳请各位专家和广大读者不吝指教和帮助。

知识就是力量,知识是创新的源泉。

史忠植

2011年1月

目 录

Contents

第 1 章 绪论	1
1.1 知识	1
1.2 知识发现的过程	3
1.3 知识发现的任务	5
1.4 知识发现的方法	8
1.4.1 统计方法	8
1.4.2 机器学习	10
1.4.3 神经计算	13
1.4.4 可视化	14
1.5 知识发现的对象	15
1.5.1 数据库	15
1.5.2 文本	16
1.5.3 Web 信息	17
1.5.4 空间数据	18
1.5.5 图像和视频数据	19
1.6 知识发现系统	19
第 2 章 决策树	25
2.1 归纳学习	25

2.2	决策树学习	26
2.3	CLS 学习算法	29
2.4	ID3 学习算法	30
2.4.1	信息论简介	30
2.4.2	信息论在决策树学习中的意义及应用	30
2.4.3	ID3 算法	31
2.4.4	ID3 算法应用举例	32
2.4.5	C4.5 算法	34
2.5	决策树的改进算法	35
2.5.1	二叉树判定算法	35
2.5.2	按信息比值进行估计的方法	36
2.5.3	按分类信息估值	37
2.5.4	按划分距离估值的方法	37
2.6	决策树的评价	38
2.7	简化决策树	40
2.7.1	简化决策树的动机	41
2.7.2	决策树过大的原因	41
2.7.3	控制树的大小	43
2.7.4	修改测试属性空间	45
2.7.5	改进测试属性选择方法	48
2.7.6	对数据进行限制	50
2.7.7	改变数据结构	51
2.8	连续性属性离散化	55
2.9	基于偏置变换的决策树学习算法 BSDT	56
2.9.1	偏置的形式化	56
2.9.2	表示偏置变换	58
2.9.3	算法描述	59
2.9.4	过程偏置变换	60
2.9.5	基于偏置变换的决策树学习算法 BSDT	63
2.9.6	经典案例库维护算法 TCBM	63
2.9.7	偏置特征抽取算法	64
2.9.8	改进的决策树生成算法 GSD	65
2.9.9	实验结果	67
2.10	单变量决策树的并行处理	68
2.10.1	并行决策树算法	68

2.10.2 串行算法的并行化	71
2.11 归纳学习中的问题	73
第3章 支持向量机	74
3.1 统计学习问题	74
3.1.1 经验风险	74
3.1.2 VC 维	75
3.2 学习过程的一致性	75
3.2.1 学习过程一致性的经典定义	75
3.2.2 学习理论的重要定理	76
3.2.3 VC 熵	76
3.3 结构风险最小归纳原理	77
3.4 支持向量机	80
3.4.1 线性可分	80
3.4.2 线性不可分	82
3.5 核函数	83
3.5.1 多项式核函数	83
3.5.2 径向基函数	84
3.5.3 多层感知机	84
3.5.4 动态核函数	84
3.6 邻近支持向量机	85
3.7 极端支持向量机	88
第4章 迁移学习	93
4.1 概述	93
4.2 相似性关系	94
4.2.1 语义相似性	95
4.2.2 结构相似性	96
4.2.3 样本相似性	96
4.2.4 相似性计算	97
4.3 归纳迁移学习	98
4.3.1 基于采样的归纳迁移	98
4.3.2 基于特征的归纳迁移	99
4.3.3 基于参数的归纳迁移	100
4.4 推导迁移学习	100

4.4.1	基于采样的知识迁移	100
4.4.2	基于特征的知识迁移	101
4.5	主动迁移学习	101
4.5.1	主动学习	101
4.5.2	主动迁移学习算法	103
4.5.3	迁移学习分类器	104
4.5.4	决策函数	105
4.6	多源领域知识的迁移学习	106
4.7	强化学习中的迁移	107
4.7.1	行为迁移	107
4.7.2	知识迁移	109
第5章	聚类分析	111
5.1	概述	111
5.2	相似性度量	112
5.2.1	相似系数	112
5.2.2	属性的相似度量	115
5.3	划分方法	116
5.3.1	k 均值算法	116
5.3.2	k 中心点算法	117
5.3.3	大型数据库的划分方法	117
5.4	层次聚类方法	119
5.4.1	BIRCH 算法	120
5.4.2	CURE 算法	120
5.4.3	ROCK 算法	121
5.5	基于密度的聚类	122
5.6	基于网格方法	125
5.7	基于模型方法	127
5.8	模糊聚类	129
5.8.1	传递闭包法	129
5.8.2	动态直接聚类法	129
5.8.3	最大树法	130
5.9	蚁群聚类方法	132
5.9.1	基本模型	132
5.9.2	LF 算法	133

5.9.3 基于群体智能的聚类算法 CSI	134
5.9.4 混合聚类算法 CSIM	136
5.10 聚类方法的评价	137
第 6 章 关联规则	140
6.1 概述	140
6.2 基本概念	141
6.3 二值型关联规则挖掘	143
6.3.1 AIS 算法	143
6.3.2 SETM 算法	144
6.3.3 Apriori 算法	146
6.3.4 Apriori 算法的改进	148
6.4 频繁模式树挖掘算法	149
6.5 垂直挖掘算法	152
6.6 挖掘关联规则的数组方法	155
6.7 频繁闭项集的挖掘算法	157
6.8 最大频繁项集挖掘算法	159
6.9 增量式关联规则挖掘	163
6.10 模糊关联规则的挖掘	166
6.11 任意多表间关联规则的并行挖掘	169
6.11.1 问题的形式描述	169
6.11.2 单表内大项集的并行计算	170
6.11.3 任意多表间大项集的生成	171
6.11.4 跨表间关联规则的提取	172
6.12 基于分布式系统的关联规则挖掘算法	173
6.12.1 候选集的生成	174
6.12.2 候选数据集的本地剪枝	175
6.12.3 候选数据集的全局剪枝	178
6.12.4 合计数轮流检测	179
6.12.5 分布式挖掘关联规则的算法	180
第 7 章 粗糙集	184
7.1 概述	184
7.1.1 知识的分类观点	186
7.1.2 新型的隶属关系	187

7.1.3	概念的边界观点	188
7.2	知识的约简	189
7.2.1	一般约简	189
7.2.2	相对约简	190
7.2.3	知识的依赖性	191
7.3	决策表的约简	192
7.3.1	属性的依赖性	192
7.3.2	一致决策表的约简	192
7.3.3	非一致决策表的约简	199
7.4	粗糙集的扩展模型	203
7.4.1	可变精度粗糙集模型	204
7.4.2	相似模型	205
7.4.3	基于粗糙集的非单调逻辑	205
7.4.4	与其他数学工具的结合	206
7.5	粗糙集的实验系统	206
7.6	粒度计算	208
7.6.1	模糊集模型	209
7.6.2	粗糙集模型	210
7.6.3	商空间理论模型	210
7.6.4	相容粒度空间模型	211
第8章	神经网络	215
8.1	概述	215
8.1.1	基本的神经网络模型	215
8.1.2	神经网络的学习方法	216
8.2	人工神经元及感知机模型	217
8.2.1	基本神经元	217
8.2.2	感知机模型	219
8.3	前向神经网络	220
8.3.1	前向神经网络模型	220
8.3.2	多层前向神经网络的误差反向传播(BP)算法	221
8.3.3	BP算法的若干改进	224
8.4	径向基函数神经网络	228
8.4.1	插值问题	229
8.4.2	正则化问题	230

8.4.3	RBF 网络学习方法	232
8.5	反馈神经网络	235
8.5.1	离散型 Hopfield 网络	235
8.5.2	连续型 Hopfield 网络	243
8.5.3	Hopfield 网络应用	245
8.5.4	双向联想记忆模型	245
8.6	随机神经网络	247
8.6.1	模拟退火算法	247
8.6.2	玻尔兹曼机	250
8.7	自组织特征映射神经网络	253
8.7.1	网络的拓扑结构	253
8.7.2	网络自组织算法	254
8.7.3	监督学习	255
第 9 章	贝叶斯网络	256
9.1	概述	256
9.1.1	贝叶斯网络的发展历史	256
9.1.2	贝叶斯方法的基本观点	257
9.1.3	贝叶斯网络在数据挖掘中的应用	258
9.2	贝叶斯概率基础	260
9.2.1	概率论基础	260
9.2.2	贝叶斯概率	263
9.3	贝叶斯学习理论	265
9.3.1	几种常用的先验分布选取方法	266
9.3.2	计算学习机制	269
9.3.3	贝叶斯问题求解	270
9.4	简单贝叶斯学习模型	273
9.4.1	简单贝叶斯模型	273
9.4.2	简单贝叶斯模型的提升	275
9.4.3	提升简单贝叶斯分类的计算复杂性	277
9.5	贝叶斯网络的建造	278
9.5.1	贝叶斯网络的结构及建立方法	278
9.5.2	学习贝叶斯网络的概率分布	279
9.5.3	学习贝叶斯网络的网络结构	281
9.6	贝叶斯潜在语义模型	284

9.7 半监督文本挖掘算法	288
9.7.1 网页聚类	288
9.7.2 对含有潜在类别主题词的文档的类别标注	289
9.7.3 基于简单贝叶斯模型学习标注和未标注样本	290
第 10 章 隐马尔可夫模型	295
10.1 马尔可夫过程	295
10.2 隐马尔可夫模型	296
10.3 评估问题	299
10.3.1 前向算法	299
10.3.2 后向算法	300
10.4 Viterbi 算法	301
10.5 学习算法	303
10.6 嵌入式隐马尔可夫模型	305
10.7 基于状态驻留时间的分段概率模型	308
第 11 章 图挖掘	312
11.1 概述	312
11.2 图的基础知识	315
11.2.1 图同构	316
11.2.2 频繁子图	317
11.3 频繁子图挖掘	317
11.3.1 基于 Apriori 的算法	317
11.3.2 基于模式增长的算法	319
11.4 约束图模式挖掘	322
11.4.1 特殊的子图挖掘	322
11.4.2 基于约束的子结构模式挖掘	322
11.5 图分类	323
11.5.1 基于核的图分类方法	323
11.5.2 最优核矩阵学习	324
11.5.3 组合维核方法	324
11.6 图模型	327
11.7 图像标注模型	333
11.7.1 混合生成式和判别式模型的图像语义标注框架	333
11.7.2 构造集群分类器链	334

11.8 社会网络分析	337
11.8.1 中心度分析	337
11.8.2 子群分析	339
11.8.3 社会网络分析的应用	341
11.8.4 社会网络分析软件	342
第 12 章 进化计算	346
12.1 概述	346
12.2 进化系统理论的形式模型	348
12.3 达尔文进化算法	350
12.4 基本遗传算法	351
12.4.1 基本遗传算法的构成要素	351
12.4.2 基本遗传算法的一般框架	352
12.5 遗传算法的数学理论	355
12.5.1 模式定理	355
12.5.2 积木块假设	358
12.5.3 隐并行性	359
12.6 遗传算法编码方法	360
12.6.1 二进制编码方法	361
12.6.2 格雷码编码方法	361
12.6.3 浮点数编码方法	362
12.6.4 符号编码方法	363
12.6.5 多参数级联编码方法	363
12.6.6 多参数杂交编码方法	363
12.7 适应度函数	364
12.8 遗传操作	366
12.8.1 选择算子	366
12.8.2 杂交算子	369
12.8.3 变异算子	371
12.8.4 反转操作	372
12.9 变长度染色体遗传算法	372
12.10 小生境遗传算法	373
12.11 混合遗传算法	374
12.12 并行遗传算法	376
12.13 分类器系统	378